Machine Recognition of Human Activities: A Survey

Article *in* IEEE Transactions on Circuits and Systems for Video Technology · December 2008

DOI: 10.1109/TCSVT.2008.2005594 · Source: IEEE Xplore

4 authors, including:

Pavan K. Turaga
Arizona State University
244 PUBLICATIONS **6,665** CITATIONS

SEE PROFILE

Rama Chellappa
University of Maryland, College Park
1,014 PUBLICATIONS **73,437** CITATIONS

SEE PROFILE

Machine Recognition of Human Activities: A Survey

Pavan Turaga, *Student Member, IEEE*, Rama Chellappa, *Fellow, IEEE*, V. S. Subrahmanian, and Octavian Udrea

***Abstract*—The past decade has witnessed a rapid proliferation of video cameras in all walks of life and has resulted in a tremendous explosion of video content. Several applications such as content-based video annotation and retrieval, highlight extraction and video summarization require recognition of the activities occurring in the video. The analysis of human activities in videos is an area with increasingly important consequences from security and surveillance to entertainment and personal archiving. Several challenges at various levels of processing—robustness against errors in low-level processing, view and rate-invariant representations at midlevel processing and semantic representation of human activities at higher level processing—make this problem hard to solve. In this review paper, we present a comprehensive survey of efforts in the past couple of decades to address the problems of representation, recognition, and learning of human activities from video and related applications. We discuss the problem at two major levels of complexity: 1) "actions" and 2) "activities." "Actions" are characterized by simple motion patterns typically executed by a single human. "Activities" are more complex and involve coordinated actions among a small number of humans. We will discuss several approaches and classify them according to their ability to handle varying degrees of complexity as interpreted above. We begin with a discussion of approaches to model the simplest of action classes known as atomic or primitive actions that do not require sophisticated dynamical modeling. Then, methods to model actions with more complex dynamics are discussed. The discussion then leads naturally to methods for higher level representation of complex activities.**

***Index Terms*—Human activity analysis, image sequence analysis, machine vision, surveillance.**

I. Introduction

RECOGNIZING human activities from video is one of the most promising applications of computer vision. In recent years, this problem has caught the attention of researchers from industry, academia, security agencies, consumer agencies, and the general populace as well. One of the earliest investigations into the nature of human motion was conducted by the contemporary photographers E. J. Marey and E. Muybridge in the 1850s who photographed moving subjects and revealed several interesting and artistic aspects involved in human and animal locomotion. The classic moving light display (MLD) experiment of Johansson [1] provided a great impetus to the study and analysis of human motion perception in the field of neuroscience. This then paved the way for mathematical modeling of human action and automatic recognition, which naturally fall into the purview of computer vision and pattern recognition.

To state the problem in simple terms, given a sequence of images with one or more persons performing an activity, can a system be designed that can automatically recognize what activity is being or was performed? As simple as the question seems, the solution has been that much harder to find. In this survey paper, we review the major approaches that have been pursued over the last 20 years to address this problem.

Several related survey papers have appeared over the years. Most notable among them are the following. Aggarwal and Cai [2] discuss three important subproblems that together form a complete action recognition system—extraction of human body structure from images, tracking across frames, and action recognition. Cedras and Shah [3] present a survey on motion-based approaches to recognition as opposed to structure-based approaches. They argue that motion is a more important cue for action recognition than the structure of the human body. Gavrila [4] presented a survey focused mainly on tracking of hands and humans via 2-D or 3-D models and a discussion of action recognition techniques. More recently, Moeslund *et al.* [5] presented a survey of problems and approaches in human motion capture including human model initialization, tracking, pose estimation, and activity recognition. Since the mid 1990s, interest has shifted more toward recognizing actions from tracked motion or structure features and on recognizing complex activities in real-world settings. Hence, this survey will focus exclusively on approaches for recognition of action and activities from video and not on the lower level modules of detection and tracking, which is discussed at length in earlier surveys [2]–[6].

The terms "action" and "activity" are frequently used interchangeably in the vision literature. In the ensuing discussion, by "actions" we refer to simple motion patterns usually executed by a single person and typically lasting for short durations of time, on the order of tens of seconds. Examples of actions include bending, walking, swimming, etc. (e.g., Fig. 1). On the other hand, by "activities" we refer to the complex sequence of actions performed by several humans who could be interacting with each other in a constrained manner. They are typically characterized by much longer temporal durations, e.g., two persons shaking hands, a football team scoring a goal, or a coordinated bank attack by multiple robbers (Fig. 2). This is not a hard boundary and there is a significant "gray area" between these two extremes. For example, the gestures of a music conductor conducting an orchestra or the constrained dynamics of a group of humans (Fig. 3) is neither as simple as an "action" nor as complex as an "activity" according to the above interpretation. However, this simple categorization provides a starting point to organize the numerous approaches that have been pro-

Manuscript received February 25, 2008; revised June 19, 2008. First published September 26, 2008; current version published October 29, 2008. This work was supported in part by the U.S. Government VACE program. This paper was recommended by Associate Editor D. Xu

The authors are with the Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA (e-mail: pturaga@umiacs.umd.edu; rama@umiacs.umd.edu; vs@umiacs.umd.edu; udrea@umiacs.umd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2008.2005594

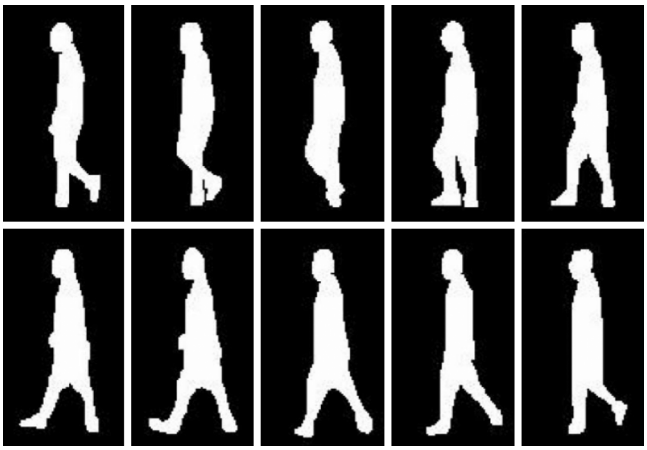


Fig. 1. Near-field video: Example of walking action. Figure taken from [7].

posed to solve the problem. A quick preview of the various approaches that fall under each of these categories is shown in Fig. 4. Real-life activity recognition systems typically follow a hierarchical approach. At the lower levels are modules such as background–foreground segmentation, tracking and object detection. At the midlevel are action–recognition modules. At the high level are the reasoning engines that encode the activity semantics based on the lower level action primitives. Thus, it is necessary to gain an understanding of both these problem domains to enable real-life deployment of systems.

The rest of this paper is organized as follows. First, we discuss a few motivating application domains in Section II. Section III provides an overview of methods for extraction of low-level image features. In Section IV, we discuss approaches for recognizing "actions." Then, in Section V, we discuss methods to represent and recognize higher level "activities." In Section VI, we discuss some open research issues for action and activity recognition and provide concluding remarks.

II. Applications

In this section, we present a few application areas that will highlight the potential impact of vision-based activity recognition systems.

1) Behavioral Biometrics: Biometrics involves study of approaches and algorithms for uniquely recognizing humans based on physical or behavioral cues. Traditional approaches are based on fingerprint, face, or iris and can be classified as physiological biometrics—i.e., they rely on physical attributes for recognition. These methods require cooperation from the subject for collection of the biometric. Recently, "behavioral biometrics" have been gaining popularity, where the premise is that behavior is as useful a cue to recognize humans as their physical attributes. The advantage of this approach is that subject cooperation is not necessary and it can proceed without interrupting or interfering with the subject's activity. Since observing behavior implies longer term observation of the subject, approaches for action recognition extend naturally to this task. Currently, the most promising example of behavioral biometrics is human gait [10].

2) Content-Based Video Analysis: Video has become a part of our everyday life. With video sharing websites experiencing relentless growth, it has become necessary to develop efficient indexing and storage schemes to improve user experience. This requires learning of patterns from raw video and summarizing a video based on its content. Content-based video summarization has been gaining renewed interest with corresponding advances in content-based image retrieval (CBIR) [11]. Summarization and retrieval of consumer content such as sports videos is one of the most commercially viable applications of this technology [12].

3) Security and Surveillance: Security and surveillance systems have traditionally relied on a network of video cameras monitored by a human operator who needs to be aware of the activity in the camera's field of view. With recent growth in the number of cameras and deployments, the efficiency and accuracy of human operators has been stretched. Hence, security agencies are seeking vision-based solutions to these tasks that can replace or assist a human operator. Automatic recognition of anomalies in a camera's field of view is one such problem that has attracted attention from vision researchers (cf., [9] and [13]). A related application involves searching for an activity of interest in a large database by learning patterns of activity from long videos [14], [15].

4) Interactive Applications and Environments: Understanding the interaction between a computer and a human remains one of the enduring challenges in designing human–computer interfaces. Visual cues are the most important mode of nonverbal communication. Effective utilization of this mode such as gestures and activity holds the promise of helping in creating computers that can better interact with humans. Similarly, interactive environments such as smart rooms [16] that can react to a user's gestures can benefit from vision-based methods. However, such technologies are still not mature enough to stand the "turing test" and thus continue to attract research interest.

5) Animation and Synthesis: The gaming and animation industry rely on synthesizing realistic humans and human motion. Motion synthesis finds wide use in the gaming industry where the requirement is to produce a large variety of motions with some compromise on the quality. The movie industry on the other hand has traditionally relied more on human animators to provide high-quality animation. However, this trend is fast changing [17]. With improvements in algorithms and hardware, much more realistic motion synthesis is now possible. A related application is learning in simulated environments. Examples of this include training of military soldiers, firefighters, and other rescue personnel in hazardous situations with simulated subjects.

III. General Overview

A generic action or activity recognition system can be viewed as proceeding from a sequence of images to a higher level interpretation in a series of steps. The major steps involved are the following:

1) input video or sequence of images;
2) extraction of concise low-level features;
3) midlevel action descriptions from low-level features;
4) high-level semantic interpretations from primitive actions.

(a) (b) (c) (d)

Fig. 2. Medium-field video: Example video sequence of a simulated bank attack (courtesy [8]). (a) Person enters the bank. (b) Robber is identified to be an outsider. Robber is entering the bank safe. (c) A customer escapes. (d) Robber makes an exit.

Fig. 3. Far-field video: Modeling dynamics of groups of humans as a deforming shape. Figure taken from [9].

In this section, we will briefly discuss some relevant aspects of item 2, i.e., low-level feature extraction. Items 3 and 4 in the list will form the subject of discussion of Sections IV and V, respectively.

Videos consist of massive amounts of raw information in the form of spatio–temporal pixel intensity variations. However, most of this information is not directly relevant to the task of understanding and identifying the activity occurring in the video. A classic experiment by Johansson [1] demonstrated that humans can perceive gait patterns from point light sources placed at a few limb joints with no additional information. Extraneous factors such as the color of the clothes, illumination conditions, background clutter do not aid in the recognition task. We briefly describe a few popular low-level features and refer the readers to other sources for a more in-depth treatment as we progress.

A. Optical Flow

Optical flow is defined as the apparent motion of individual pixels on the image plane. Optical flow often serves as a good approximation of the true physical motion projected onto the image plane. Most methods to compute optical flow assume that the color/intensity of a pixel is invariant under the displacement from one video frame to the next. We refer the reader to [18] for a comprehensive survey and comparison of optical flow computation techniques. Optical flow provides a concise description of both the regions of the image undergoing motion and the velocity of motion. In practice, computation of optical flow is susceptible to noise and illumination changes. Applications include [19], which used optical flow to detect and track vehicles in an automated traffic surveillance application.

B. Point Trajectories

Trajectories of moving objects have popularly been used as features to infer the activity of the object (see Fig. 5). The image-plane trajectory itself is not very useful as it is sensitive to translations, rotations, and scale changes. Alternative representations such as trajectory velocities, trajectory speeds, spatio–temporal curvature, relative motion, etc., have been proposed that are invariant to some of these variabilities. A good survey of these approaches can be found in [3]. Extracting unambiguous point trajectories from video is complicated by several factors such as occlusions, noise, and background clutter. Accurate tracking algorithms need to be employed for obtaining motion trajectories [6].

C. Background Subtracted Blobs and Shape

Background subtraction is a popular method to isolate the moving parts of a scene by segmenting it into background and foreground (cf., [21]). As an example, from the sequence of background subtracted images shown in Fig. 1, the human's walking action can be easily perceived. The shape of the human silhouette plays a very important role in recognizing human actions, and it can be extracted from background subtraction blobs (see Fig. 6). Several methods based on global, boundary, and skeletal descriptors have been proposed to quantify shape. Global methods such as moments [22] consider the entire shape region to compute the shape descriptor. Boundary methods on the other hand consider only the shape contour as the defining characteristic of the shape. Such methods include chain codes [23] and landmark-based shape descriptors [24]. Skeletal methods represent a complex shape as a set of 1-D skeletal curves, for example, the medial axis transform [25]. Applications include shape-based dynamic modeling of the human silhouette as in [26] to perform gait recognition.

D. Filter Responses

There are several other features that can be broadly classified as based on spatio–temporal filter responses. In their work, Zhong *et al.* [13] process a video sequence using a spatial Gaussian and a derivative of Gaussian on the temporal axis. Due to the derivative operation on the temporal axis, the filter shows high responses at regions of motion. This response was then thresholded to yield a binary motion mask followed by aggregation into spatial histogram bins. Such a feature encodes motion and its corresponding spatial information compactly and is useful for far-field and medium-field surveillance videos. The notion of scale-space filtering has also been extended to videos by several researchers. Laptev *et al.* [27] propose a generalization of the Harris corner detector to videos using a set of spatio–temporal Gaussian derivative filters. Similarly, Dollar

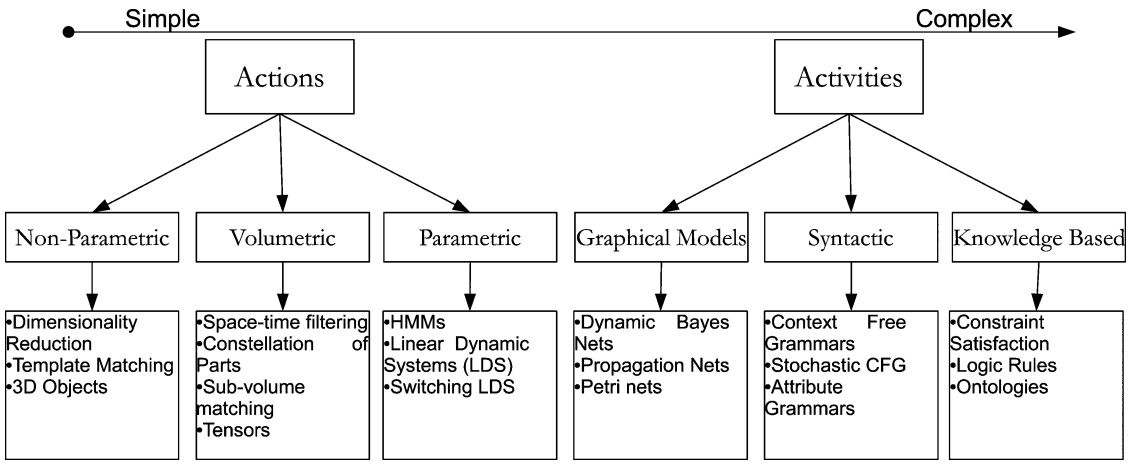


Fig. 4. Overview of approaches for action and activity recognition.

Fig. 5. Trajectories of a passenger and luggage cart. The wide difference in the trajectories is indicative of the difference in activities. Figure taken from [20].

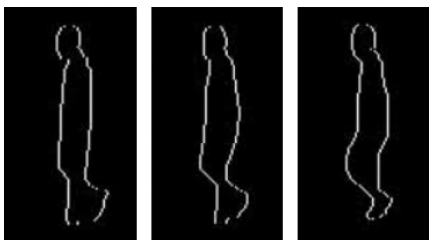


Fig. 6. Silhouettes extracted from the walking sequence shown in Fig. 1. Silhouettes encode sufficient information to recognize actions. Figure taken from [7].

et al. [28] extract distinctive periodic motion-based landmarks in a given video using a Gaussian kernel in space and a Gabor function in time. Because these approaches are based on simple convolution operations, they are fast and easy to implement. They are quite useful in scenarios with low-resolution or poor quality video where it is difficult to extract other features such as optical flow or silhouettes.

IV. Modeling and Recognizing Actions

Approaches for modeling actions can be categorized into three major classes—nonparametric, volumetric, and parametric time-series approaches. Nonparametric approaches typically extract a set of features from each frame of the video. The features are then matched to a stored template. Volumetric approaches on the other hand do not extract features on a frame-by-frame basis. Instead, they consider a video as a 3-D volume of pixel intensities and extend standard image features such as scale-space extrema, spatial filter responses, etc., to the 3-D case. Parametric time-series approaches specifically impose a model on the temporal dynamics of the motion. The particular parameters for a class of actions is then estimated from training data. Examples of parametric approaches include hidden Markov models (HMMs), linear dynamical systems (LDSs), etc. We will first discuss the nonparametric methods, then the volumetric approaches, and finally the parametric time-series methods.

A. Nonparametric Approaches for Action Recognition

1) 2-D Templates: One of the earliest attempts at action recognition without relying on 3-D structure estimation was proposed by Polana and Nelson [29]. First, they perform motion detection and tracking of humans in the scene. After tracking, a "cropped" sequence containing the human is constructed. Scale changes are compensated for by normalizing the size of the human. A periodicity index is computed for the given action and the algorithm proceeds to recognize the action if it is found to be sufficiently periodic. To perform recognition, the periodic sequence is segmented into individual cycles using the periodicity estimate and combined to get an average cycle. The average cycle is divided into a few temporal segments and flow-based features are computed for each spatial location in each segment. The flow features in each segment are averaged into a single frame. The average-flow frames within an activity cycle form the templates for each action class.

Bobick and Davis [30] proposed "temporal templates" as models for actions. In their approach, the first step involved is background subtraction, followed by an aggregation of a sequence of background subtracted blobs into a single static image. They propose two methods of aggregation—the first method gives equal weight to all images in the sequence, which gives rise to a representation called the "motion energy image"

Fig. 7. Temporal templates similar to [30]. Left: motion energy image of a sequence of a person raising both hands. Right: motion history image of the same action.

(MEI). The second method gives decaying weights to the images in the sequence with higher weight given to new frames and low weight to older frames. This leads to a representation called the "motion history image" (MHI) (for example, see Fig. 7). The MEI and MHI together comprise a template for a given action. From the templates, translation, rotation, and scale invariant Hu moments [22] are extracted that are then used for recognition. It was shown in [30] that MEI and MHI have sufficient discriminating ability for several simple action classes such as "sitting down," "bending," "crouching," and other aerobic postures. However, it was noted in [31] that MEI and MHI lose discriminative power for complex activities due to overwriting of the motion history and hence are unreliable for matching.

2) 3-D Object Models: Successful application of models and algorithms to object recognition problems led researchers in action recognition to propose alternate representations of actions as spatio–temporal objects. Syeda-Mahmood *et al.* proposed a representation of actions as generalized cylinders in the joint (x, y, t) space [32]. Yilmaz and Shah [33] represent actions as 3-D objects induced by stacking together tracked 2-D object contours. A sequence of 2-D contours in (x, y) space can be treated as an object in the joint (x, y, t) space. This representation encodes both the shape and motion characteristics of the human. From the (x, y, t) representation, concise descriptors of the object's surface are extracted corresponding to geometric features such as peaks, pits, valleys, and ridges. Because this approach is based on stacking together a sequence of silhouettes, accurate correspondence between points of successive silhouettes in the sequences needs to be established. Quasi-view invariance for this representation was shown theoretically by assuming an affine camera model. Similar to this approach, Gorelick *et al.* [34] proposed using background subtracted blobs instead of contours, which are then stacked together to create an (x, y, t) binary space–time (ST) volume (for example, see Fig. 8). Because this approach uses background subtracted blobs, the problem of establishing correspondence between points on contours in the sequence does not exist. From this ST volume, 3-D shape descriptors are extracted by solving a Poisson equation [34]. Because these approaches require careful segmentation of background and the foreground, they are limited in applicability to fixed camera settings.

3) Manifold Learning Methods: Most approaches in action recognition involve dealing with data in very high-dimensional spaces. Hence, these approaches often suffer from the "curse of dimensionality." The feature space becomes sparser in an exponential fashion with the dimension, thus requiring a larger

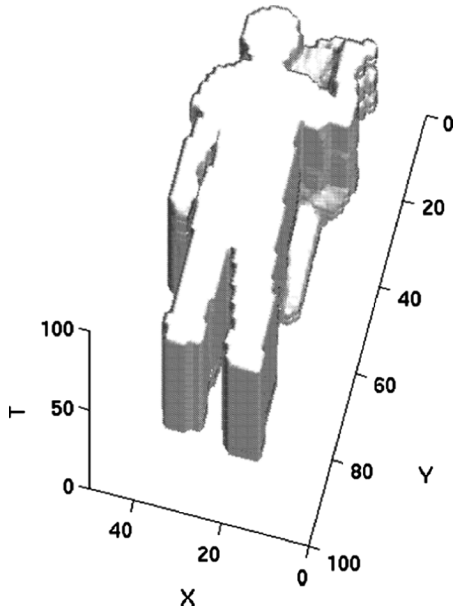


Fig. 8. The 3-D space–time object, similar to [34], obtained by stacking together binary background subtracted images of a person waving his hand.

number of samples to build efficient class-conditional models. Learning the manifold on which the data resides enables us to determine the inherent dimensionality of the data as opposed to the raw dimensionality. The inherent dimensionality contains fewer degrees of freedom and allows efficient models to be designed in the lower dimensional space. The simplest way to reduce dimensionality is via principal component analysis (PCA), which assumes that the data lies on a linear subspace. Except in very special cases, data does not lie on a linear subspace, thus requiring methods that can learn the intrinsic geometry of the manifold from a large number of samples. Nonlinear dimensionality reduction techniques allow for representation of data points based on their proximity to each other on nonlinear manifolds. Several methods for dimensionality reduction such as PCA, locally linear embedding (LLE) [35], Laplacian eigenmap [36], and Isomap [37] have been applied to reduce the high-dimensionality of video data in action–recognition tasks (cf., [38]–[40]). Specific recognition algorithms such as template matching, dynamical modeling, etc., can be performed more efficiently once the dimensionality of the data has been reduced.

B. Volumetric Approaches

1) Spatio–Temporal Filtering: These approaches are based on filtering a video volume using a large filter bank. The responses of the filter bank are further processed to derive action specific features. These approaches are inspired by the success of filter-based methods on other still image recognition tasks such as texture segmentation [41]. Further, spatio–temporal filter structures such as oriented Gaussian kernels and their derivatives [42] and oriented Gabor filter banks [43] have been hypothesized to describe the major spatio–temporal properties of cells in the visual cortex. Chomat *et al.* [44] model a segment of video as a (x, y, t) spatio–temporal volume and compute local appearance models at each pixel using a Gabor filter

bank at various orientation and spatial scales and a single temporal scale. A given action is recognized using a spatial average of the probabilities of individual pixels in a frame. Because actions are analyzed at a single temporal scale, this method is not applicable to variations in execution rate. As an extension to this approach, local histograms of normalized ST gradients at several temporal scales are extracted by Zelnik-Manor and Irani [45]. The sum of the chi-square metric between histograms is used to match an input video with a stored exemplar. Filtering with the Gaussian kernel in space and the derivative of the Gaussian on the temporal axis followed by thresholding of the responses and accumulation into spatial histograms was found to be a simple yet effective feature for actions in far-field settings [13].

Filtering approaches are fast and easy to implement due to efficient algorithms for convolution. In most applications, the appropriate bandwidth of the filters is not known *a priori*, thus a large filter bank at several spatial and temporal scales is required for effectively capturing the action dynamics. Moreover, the response generated by each filter has the same dimensions as the input volume, hence using large filter banks at several spatial and temporal scales is prohibitive.

2) Part-Based Approaches: Several approaches have been proposed that consider a video volume as a collection of local parts, where each part consists of some distinctive motion pattern. Laptev and Lindeberg [27] proposed a spatio–temporal generalization of the well-known Harris interest point detector, which is widely used in object recognition applications and applied it to modeling and recognizing actions in ST. This method is based on the 3-D generalization of scale-space representations. A given video is convolved with a 3-D Gaussian kernel at various spatial and temporal scales. Then, spatio–temporal gradients are computed at each level of the scale-space representation. These are then combined within a neighborhood of each point to yield stable estimates of the spatio–temporal second-moment matrix. Local features are then derived from these smoothed estimates of gradient moment matrices. In a similar approach, Dollar *et al.* [28] model a video sequence by the distribution of ST feature prototypes. The feature prototypes are obtained by k-means clustering of a large set of features—ST gradients—extracted at ST interest points from the training data. Neibles *et al.* [46] use a similar approach where they use a bag-of-words model to represent actions. The bag-of-words model is learned by extracting spatio–temporal interest points and clustering of the features. These interest points can be used in conjunction with machine learning approaches such as support vector machines (SVMs) [47] and graphical models [46]. Because the interest points are local in nature, longer term temporal correlations are ignored in these approaches. To address this issue, a method based on correlograms of prototype labels was presented in [48]. In a slightly different approach Nowozin *et al.* [49] consider a video as a sequence of sets, where each set consists of the parts found in a small temporally sliding window. These approaches do not directly model the global geometry of local parts instead considering them as a bag of features. Different actions may be composed of similar ST parts but may differ in their geometric relationships. Integrating global geometry into the part-based video representation was investigated by Boiman *et al.* [50] and Wong *et al.* [51]. This approach may be termed as a constellation of parts as opposed to the simpler bag-of-parts model. Computational complexity can be large for constellation models with a large number of parts, which is typically the case for human actions. Song *et al.* [52] addressed this issue by approximating the connections in the constellation via triangulation. Niebles *et al.* [53] proposed a hierarchical model where the higher level is a constellation of parts much smaller than the actual number of features. Each of the parts in the constellation consists of a bag of features at the lower level. This approach combines the advantages of both the bag of features and the constellation model and preserves computational efficiency at the same time.

In most of these approaches, the detection of the parts is usually based on linear operations such as filtering and spatio–temporal gradients, hence the descriptors are sensitive to changes in appearance, noise, occlusions, etc. It has also been noted that interest points are extremely sparse in smooth human actions and certain types of actions do not give rise to distinctive features [28], [46]. However, due to their local nature, they are more robust to nonstationary backgrounds.

3) Subvolume Matching: As opposed to part-based approaches, researchers have also investigated matching of videos by matching subvolumes between a video and a template. Shechtman *et al.* [54] present an approach derived from ST motion-based correlation to match actions with a template. The main difference of this approach from the part-based approaches is that it does not extract action descriptors from extrema in scale space, rather it looks for similarity between local ST patches based on how similar the motion is in the two patches. However, computing this correlation throughout a given video volume can be computationally intensive. Inspired by the success of Haar-type features or "box features" in object detection [55], Ke *et al.* [56] extended this framework to 3-D. In their approach, they define 3-D Haar-type features that are essentially outputs of 3-D filter banks with $+1$'s and -1's as the filter coefficients. These filter responses used in conjunction with boosting approaches result in robust performance. In another approach, Ke *et al.* [57] consider a video volume as a collection of subvolumes of arbitrary shape, where each subvolume is a spatially coherent region. The subvolumes are obtained by clustering the pixels based on appearance and spatial proximity. A given video is oversegmented into many subvolumes or "supervoxels." An action template is matched by searching among the oversegmented volumetric regions and finding the minimal set of regions that maximize overlap between their union and the template.

Subvolume matching approaches such as these are susceptible to changing backgrounds but are more robust to noise and occlusions. Another advantage is that these approaches can be extended to features such as optical flow as in [56] to achieve robustness to changes in appearance.

4) Tensor-Based Approaches: Tensors are generalizations of matrices to multiple dimensions. A 3-D ST volume can naturally be considered as a tensor with three independent dimensions. Vasilescu [58] proposed the modeling of human action, human identity, and joint angle trajectories by considering them as independent dimensions of a tensor. By decomposing the

overall data tensor into dominant modes (as a generalization of PCA), one can extract signatures corresponding to both the action and the identity of the person performing the action. Recently, Kim *et al.* [59] extended canonical correlation analysis to tensors to match videos directly to templates. In their approach, the dimensions of the tensor were simply the ST dimensions corresponding to (x, y, t). Similarly, Wolf *et al.* [60] extended low-rank SVM techniques to the space of tensors for action recognition.

Tensor-based approaches offer a direct method for holistic matching of videos without recourse to midlevel representations such as the previous ones. Moreover, they can incorporate other types of features such as optical flow, ST filter responses, etc., into the same framework by simply adding more independent dimensions to the tensor.

C. Parametric Methods

The previous section focused on representations and models that are well suited for simple actions. The parametric approaches that we will describe in this section are better suited for more complex actions that are temporally extended. Examples of such complex actions include the steps in a ballet dancing video, a juggler juggling a ball, and a music conductor conducting an orchestra using complex hand gestures.

1) Hidden Markov Models: One of the most popular state–space models is the hidden Markov model. In the discrete HMM formalism, the state space is considered to be a finite set of discrete points. The temporal evolution is modeled as a sequence of probabilistic jumps from one discrete state to the other. HMMs first found wide applicability in speech recognition applications in the early 1980s. An excellent source for a detailed explanation of HMMs and its associated three problems—inference, decoding, and learning—can be found in [61]. Beginning in the early 1990s, HMMs began to find wide applicability in computer vision systems. One of the earliest approaches to recognize human actions via HMMs was proposed by Yamato *et al.* [62] where they recognized tennis shots such as backhand stroke, backhand volley, forehand stroke, forehand volley, smash, etc., by modeling a sequence of background subtracted images as outputs of class-specific HMMs. Several successful gesture recognition systems such as in [63]–[65] make extensive use of HMMs by modeling a sequence of tracked features such as hand blobs as HMM outputs.

HMMs have also found applicability in modeling the temporal evolution of human gait patterns both for action recognition and biometrics (cf., [66] and [67]). All these approaches are based on the assumption that the feature sequence being modeled is a result of a single person performing an action. Hence, they are not effective in applications where there are multiple agents performing an action or interacting with each other. To address this issue, Brand *et al.* [68] proposed a coupled HMM to represent the dynamics of interacting targets. They demonstrate the superiority of their approach over conventional HMMs in recognizing two-handed gestures. Incorporating domain knowledge into the HMM formalism has been investigated by several researchers. Moore *et al.* [69] used HMMs in conjunction with object detection modules to exploit the relationship between actions and objects. Hongeng and Nevatia [70] incorporate *a priori* beliefs of state duration into the HMM framework and the resultant model is called hidden semi-Markov model (semi-HMMs). Cuntoor and Chellappa [71] have proposed a mixed-state HMM formalism to model nonstationary activities, where the state space is augmented with a discrete label for higher level behavior modeling.

HMMs are efficient for modeling time-sequence data and are useful both for their generative and discriminative capabilities. HMMs are well suited for tasks that require recursive probabilistic estimates [63] or when accurate start and end times for action units are unknown. However, their utility is restricted due to the simplifying assumptions that the model is based on. Most significantly the assumption of Markovian dynamics and the time-invariant nature of the model restricts the applicability of HMMs to relatively simple and *stationary* temporal patterns.

2) Linear Dynamical Systems: Linear dynamical systems are a more general form of HMMs where the state space is not constrained to be a finite set of symbols but can take on continuous values in $\mathbb{R}^k$ where k is the dimensionality of the state space. The simplest form of LDS is the first-order time-invariant Gauss–Markov processes, which is described by

$$x(t) = Ax(t-1) + w(t), \qquad w \sim N(0, Q) \tag{1}$$

$$y(t) = Cx(t) + v(t), \qquad v \sim N(0, R) \tag{2}$$

where $x \in \mathbb{R}^d$ is the d-dimensional state vector and $y \in \mathbb{R}^n$ is the n-dimensional observation vector with $d \ll n$. w and v are the process and observation noise, respectively, which are Gaussian distributed with zero-means and covariance matrices Q and R, respectively. The LDS can be interpreted as a continuous state–space generalization of HMMs with a Gaussian observation model. Several applications such as recognition of humans and actions based on gait [7], [72], [73], activity recognition [9], [74], and dynamic texture modeling and recognition [75], [76] have been proposed using LDSs.

Advances in system identification theory for learning LDS model parameters from data [77]–[79] and distance metrics on the LDS space [75], [80], [81] have made LDSs popular for learning and recognition of high-dimensional time-series data. More recently, in-depth study of the LDS space has enabled the application of machine learning tools on that space such as dynamic boosting [82], kernel methods [83], [84], and statistical modeling [85]. Newer methods to learn the model parameters [86] have made learning much more efficient than in the case of HMMs. Like HMMs, LDSs are also based on assumptions of Markovian dynamics and conditionally independent observations. Thus, as in the case of HMMs, the time-invariant model is not applicable to nonstationary actions.

3) Nonlinear Dynamical Systems: While time-invariant HMMs and LDSs are efficient modeling and learning tools, they are restricted to linear and stationary dynamics. Consider the following activity: a person bends down to pick up an object, then he walks to a nearby table and places the object on the table, and finally rests on a chair. This activity is composed of a sequence of short segments each of which can be modeled as an LDS. The entire process can be seen as switching between

LDSs. The most general form of the time-varying LDS is given by

$$x(t) = A(t)x(t-1) + w(t), \qquad w \sim N(0, Q) \tag{3}$$
$$y(t) = C(t)x(t) + v(t), \qquad v \sim N(0, R) \tag{4}$$

which looks similar to the LDS in (1) and (2), except that the model parameters A and C are allowed to vary with time. To tackle such complex dynamics, a popular approach is to model the process using switching linear dynamical systems (SLDSs) or jump linear systems (JLSs). An SLDS consists of a set of LDSs with a switching function that causes model parameters to change by switching between models. Bregler [87] presented a multilayered approach to recognize complex movements consisting of several levels of abstraction. The lowest level is a sequence of input images. The next level consists of "blob" hypotheses where each blob is a region of coherent motion. At the third level, blob tracks are grouped temporally. The final level consists of an HMM for representing the complex behavior. North *et al.* [88] augment the continuous state vector with a discrete state component to form a "mixed" state. The discrete component represents a mode of motion or more generally a "switch" state. Corresponding to each switch state, a Gaussian autoregressive model is used to represent the dynamics. A maximum-likelihood approach is used to learn the model parameters for each motion class. Pavlovic and Rehg [89], [90] model the nonlinearity in human motion in a similar framework, where the dynamics are modeled using LDS and the switching process is modeled using a probabilistic finite state machine.

Though the SLDS framework has greater modeling and descriptive power than HMMs and LDSs, learning and inference in SLDS are much more complicated, often requiring approximate methods [91]. In practice, determining the appropriate number of switching states is challenging and often requires large amounts of training data or extensive hand tuning.

V. Modeling and Recognizing Activities

Most activities of interest in applications such as surveillance and content-based indexing involve several actors, who interact not only with each other, but also with contextual entities. The approaches discussed so far are mostly concerned with modeling and recognizing actions of a single actor. Modeling a complex scene, the inherent structure and semantics of complex activities require higher level representation and reasoning methods.

A. Graphical Models

1) Belief Networks: A Bayesian network (BN) [92] is a graphical model that encodes complex conditional dependencies between a set of random variables that are encoded as local conditional probability densities (CPD). Dynamic belief networks (DBNs) are a generalization of the simpler BNs by incorporating temporal dependencies between random variables. DBNs encode more complex conditional dependence relations among several random variables as opposed to just one hidden variable as in a traditional HMM.

Huang *et al.* [19] used DBNs for vision-based traffic monitoring. Buxton and Gong [93] used BNs to capture the dependencies between scene layout and low-level image measurements for a traffic surveillance application. Remagnino *et al.* [94] present an approach using DBNs for scene description at two levels of abstraction—agent level descriptions and interagent interactions. Modeling two-person interactions such as pointing, punching, pushing, hugging, etc., was proposed by Park and Aggarwal [95] in a two-stage process. First, pose estimation is done via a BN and temporal evolution of pose is modeled by a DBN. Intille and Bobick [96] use BNs for multiagent interactions where the network structure is automatically generated from the temporal structure provided by a user. Usually the structure of the DBN is provided by a domain expert. However, this is difficult in real-life systems where there are a very large number of variables with complex interdependencies. To address this issue, Gong *et al.* [97] presented a DBN framework where the structure of the network is discovered automatically using Bayesian information criterion [98], [99].

DBNs have also been used to recognize actions using the contextual information of the objects involved. Moore *et al.* [69] conduct action recognition using belief networks based on scene context derived from other objects in the scene. Gupta *et al.* [100] present a BN for interpretation of human–object interactions that integrates information from perceptual tasks such as human motion analysis, manipulable object detection, and "object reaction" determination.

Though DBNs are more general than HMMs by considering dependencies between several random variables, the temporal model is usually Markovian as in the case of HMMs. Thus, only sequential activities can be handled by the basic DBN model. Development of efficient algorithms for learning and inference in graphical models (cf., [101]) have made them popular tools to model structured activities. Methods to learn the topology or structure of BNs from data [102] have also been investigated in the machine learning community. However, to learn the local CPDs for large networks requires very large amounts of training data or extensive hand-tuning by experts both of which limit the applicability of DBNs in large scale settings.

2) Petri Nets: Petri nets were defined by Petri [103] as a mathematical tool for describing relations between conditions and events. Petri nets are particularly useful to model and visualize behaviors such as sequencing, concurrency, synchronization, and resource sharing [104], [105]. Petri nets are bipartite graphs consisting of two types of nodes—places and transitions. Places refer to the state of an entity and transitions refer to changes in the state of the entity. Consider an example of a car pickup activity represented by a probabilistic Petri net as shown in Fig. 9. In this figure, the places are labeled $p_1, \ldots, p_5$ and transitions $t_1, \ldots, t_6$. In this PN, p_1 and p_3 are the start nodes and p_5 is the terminal node. When a car enters the scene, a "token" is placed in place p_1. The transition t_1 is enabled in this state, but it cannot fire until the condition associated with it is satisfied, i.e., when the car stops near a parking slot. When this occurs, the token is removed from p_1 and placed in p_2. Similarly, when a person enters the parking lot, a token is placed in p_3 and transition t_5 fires after the person disappears near the parked car. The token is then removed from p_3 and placed in p_4.

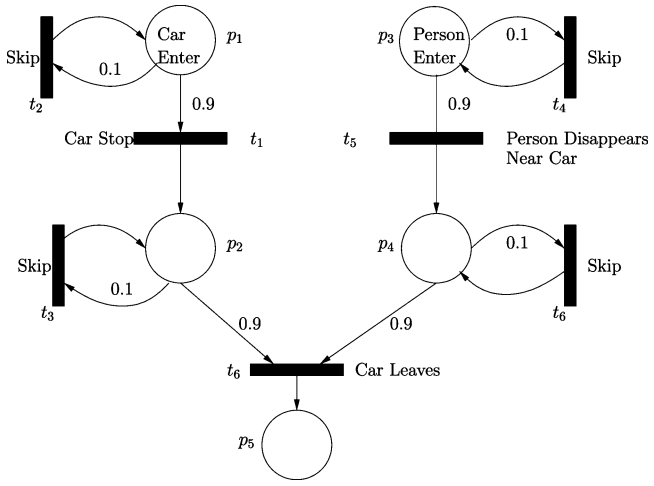


Fig. 9. Probabilistic Petri net representing a pickup-by-car activity. Figure taken from [108].

Now with a token in each of the enabling places of transition t_6, it is ready to fire when the associated condition, i.e., car leaving the parking lot is satisfied. Once the car leaves, t_6 fires and both tokens are removed and a token placed in the final place p_5. This example illustrates sequencing, concurrency, and synchronization.

Petri nets were used by Castel *et al.* [106] to develop a system for high-level interpretation of image sequences. In their approach, the structure of the Petri net was specified *a priori*. This can be tedious for large networks representing complex activities. Ghanem *et al.* [107] proposed a method to semiautomate this task by automatically mapping a small set of logical, spatial, and temporal operators to the graph structure. Using this method, they developed an interactive tool for querying surveillance videos by mapping user queries to Petri nets. However, these approaches were based on deterministic Petri nets. Hence, they cannot deal with uncertainty in the low-level modules as is usually the case with trackers, object detectors, etc. Further, real-life human activities do not conform to hard-coded models—the models need to allow deviations from the expected sequence of steps while penalizing significant deviations. To address this issue, Albanese *et al.* [108] proposed the concept of a probabilistic Petri net (PPN) (see Fig. 9). In a PPN, the transitions are associated with a weight that encodes the probability with which that transition fires. By using skip transitions and penalizing them with low probability, robustness is achieved to missing observations in the input stream. Further, the uncertainty in the identity of an object or the uncertainty in the unfolding of an activity can be efficiently incorporated into the tokens of the Petri net.

Though Petri nets are an intuitive tool for expressing complex activities, they suffer from the disadvantage of having to manually describe the model structure. The problem of learning the structure from training data has not yet been formally addressed.

3) Other Graphical Models: Other graphical models have been proposed to deal with the drawbacks in DBNs—most significantly, the limitation to sequential activities. Graphical models that specifically model more complex temporal relations such as sequentiality, duration, parallelism, synchrony, etc., have been proposed in the DBN framework. Examples include the work of Pinhanez and Bobick [109] who use a simplified version of Allen's interval algebra to model sophisticated temporal ordering constraints such as past, now, and future. This structure is termed the past–now–future (PNF) network. Similarly, Shi *et al.* [110], [111] have proposed using propagation nets to represent activities using partially ordered temporal intervals. In their approach, an activity is constrained by temporal and logical ordering and duration of the activity intervals. More recently, Hamid *et al.* [112] considered a temporally extended activity as a sequence of event labels. Due to contextual and activity specific constraints, the sequence labels are observed to have some inherent partial ordering. For example, in a kitchen setting, the refrigerator would have to be opened before the eggs can be accessed. Using these constraints, they consider an activity model as a set of subsequences, which encode the partial ordering constraints of varying lengths. These subsequences are efficiently represented using Suffix trees. The advantage of the Suffix-tree representation is that the structure of the activity can be learned from training data using standard graph-theoretic methods.

B. Syntactic Approaches

1) Grammars: Grammars express the structure of a process using a set of production rules. To draw a parallel to grammars in language modeling, production rules specify how sentences (activities) can be constructed from words (activity primitives), and how to recognize if a sentence (video) conforms to the rules of a given grammar (activity model). One of the earliest use of grammars for visual activity recognition was proposed by Brand [113], who used a grammar to recognize hand manipulations in sequences containing disassembly tasks. He made use of simple grammars with no probabilistic modeling. Ryoo and Aggarwal [114] used the context-free grammar (CFG) formalism to model and recognize composite human activities and multiperson interactions. They followed a hierarchical approach where the lower levels are composed of HMMs and BNs. The higher level interactions are modeled by CFGs. Context-free grammar approaches present a sound theoretical basis for modeling structured processes. In syntactic approaches, one only needs to enumerate the list of primitive events that need to be detected and the set of production rules that define higher level activities of interest. Once the rules of a CFG have been formulated, efficient algorithms to parse them exist [115], [116], which have made them popular in real-time applications.

Because deterministic grammars expect perfect accuracy in the lower levels, they are not suited to deal with errors in low-level tasks such as tracking errors and missing observations. In complex scenarios involving several agents requiring temporal relations that are more complex than just sequencing, such as parallelism, overlap, synchrony, it is difficult to formulate the grammatical rules manually. Learning the rules of the grammar from training data is a promising alternative, but it has proved to be extremely difficult in the general case [117].

2) Stochastic Grammars: Algorithms for detection of low-level primitives are frequently probabilistic in nature. Thus, stochastic context-free grammars (SCFGs), which are a probabilistic extension of CFGs, were found to be suitable

$$S \rightarrow BOARDING_N$$
$$BOARDING \rightarrow appear_0 CHECK_1 disappear_1$$
$$(isPerson(appear.class) \wedge isInside(appear.loc, Gate) \wedge isInside(disappear.loc, Plane))$$
$$CHECK \rightarrow moveclose_0 CHECK_1$$
$$CHECK \rightarrow moveaway_0 CHECK_1$$
$$CHECK \rightarrow moveclose_0 moveaway_1 CHECK_1$$
$$(isPerson(moveclose.class) \wedge moveclose.idr = moveaway.idr)$$

Fig. 10. Example of an attribute grammar for a passenger boarding an airplane taken from [120].

for integration with real-life vision modules. SCFGs were used by Ivanov and Bobick [118] to model the semantics of activities whose structure was assumed to be known. They used HMMs for low-level primitive detection. The grammar production rules were augmented with probabilities and a "skip" transition was introduced. This resulted in increased robustness to insertion errors in the input stream and also to errors in low-level modules. Moore *et al.* [119] used SCFGs to model multitasked activities—activities that have several independent threads of execution with intermittent dependent interactions with each other as demonstrated in a blackjack game with several participants.

In many cases, it is desirable to associate additional attributes or features to the primitive events. For example, the exact location in which the primitive event occurs may be significant for describing an event, but this may not be effectively encoded in the (finite) primitive event set. Thus, attribute grammars achieve greater expressive power than traditional grammars. Probabilistic attribute grammars have been used by Joo and Chellappa [120] for multiagent activities in surveillance settings. In the example shown in Fig. 10, one can see the production rules and the primitive events such as "appear," "disappear," "moveclose," "moveaway," etc., in the description of the activity. The primitive events are further associated with attributes such as location (loc) where the appearance and disappearance events occur, classification (class) into a set of objects, identity (idr) of the entity involved, etc.

While SCFGs are more robust than CFGs to errors and missed detections in the input stream, they share many of the temporal relation modeling limitations of CFGs as discussed above.

C. Knowledge and Logic-Based Approaches

1) Logic-Based Approaches: Logic-based methods rely on formal logical rules to describe common sense domain knowledge to describe activities. Logical rules are useful to express domain knowledge as input by a user or to present the results of high-level reasoning in an intuitive and human-readable format. Declarative models [121] describe all expected activities in terms of scene structure, events, etc. The model for an activity consists of the interactions between the objects of the scene. Medioni *et al.* [122] propose a hierarchical representation to recognize a series of actions performed by a single agent. Symbolic descriptors of actions are extracted from low-level features through several midlevel layers. Next, a rule-based method is used to approximate the probability of occurrence of a specific activity by matching the properties of the agent with the expected distributions (represented by a mean and a variance) for a particular action. In a later work, Hongeng *et al.* [123] extended this representation by considering an activity to be composed of several action threads. Each action thread is modeled as a stochastic finite state automaton. Constraints between the various threads are propagated in a temporal logic network. Shet *et al.* [124] propose a system that relies on logic programming to represent and recognize high-level activities. Low-level modules are used to detect primitive events. The high-level reasoning engine is based on Prolog and recognizes activities, which are represented by logical rules between primitives. These approaches do not explicitly address the problem of uncertainty in the observation input stream. To address this issue, a combination of logical and probabilistic models was presented in [125], where each logical rule is represented as first-order logic formula. Each rule is further provided with a weight, where the weight indicates a belief in the accuracy of the rule. Inference is performed using a Markov-logic network.

While logic-based methods are a natural way of incorporating domain knowledge, they often involve expensive constraint satisfaction checks. Further, it is not clear how much domain knowledge should be incorporated in a given setting—incorporating more knowledge can potentially make the model rigid and nongeneralizable to other settings. Further, the logic rules require extensive enumeration by a domain expert for every deployment.

2) Ontologies: In most practical deployments that use any of the aforementioned approaches, symbolic activity definitions are constructed in an empirical manner, for example, the rules of a grammar or a set of logical rules are specified manually. Though empirical constructs are fast to design and even work very well in most cases, they are limited in their utility to specific deployments for which they have been designed. Hence, there is a need for a centralized representation of activity definitions or ontologies for activities that are independent of algorithmic choices. Ontologies standardize activity definitions, allow for easy portability to specific deployments, enable interoperability of different systems, and allow easy replication and comparison of system performance. Several researchers have proposed ontologies for specific domains of visual surveillance. For example, Chen *et al.* [126] proposed an ontology for analyzing social interaction in nursing homes, Hakeem *et al.* for classification of meeting videos [127], and Georis *et al.* [8] for activities in a bank monitoring setting. To consolidate these efforts and to build a common knowledge base of domain ontologies, the Video Event Challenge Workshop was held in 2003. As a result of this workshop, ontologies have been defined for six domains of video surveillance [128]: 1) perimeter and internal security; 2) railroad crossing surveillance; 3) visual bank monitoring; 4) visual metro monitoring; 5) store security; and 6) airport-tarmac security. An example from the ontology output is shown in Fig. 11, which describes

```
PROCESS(cruise-parking-lot(vehicle v, parking-lot lot),
Sequence(enter(v, lot),
  set-to-zero(i),
  Repeat-Until(
    AND(move-in-circuit(v), inside(v, lot), increment(i)),
    equal(i, n)),
  exit(v, lot)))
```

Fig. 11. Ontology for car cruising in parking lot activity. Example taken from [128].

car cruising activity. This ontology keeps track of the number of times the car moves around in a circuit inside the parking lot without stopping. When this exceeds a set threshold, a cruising activity is detected. The workshop also led to the development of two formal languages—the video event representation language (VERL) [129], [130], which provides an ontological representation of complex events in terms of simpler subevents, and the video event markup language (VEML), which is used to annotate VERL events in videos.

Though ontologies provide concise high-level definitions of activities, they do not necessarily suggest the right "hardware" to "parse" the ontologies for recognition tasks.

VI. Directions for Future Work and Conclusion

A lot of enthusiasm has been generated in the vision community by recent advances in machine recognition of activities. However, several important issues remain to be addressed. In this section, we briefly discuss some of these issues.

A. Real-World Conditions

Most action and activity recognition systems are currently designed and tested on video sequences acquired in constrained conditions. Factors that can severely limit the applicability in real-world conditions include noise, occlusions, shadows, etc. Errors in feature extraction can easily propagate to higher levels. For real-world deployment, action recognition systems need to be tested against such real-world conditions. Methods that are robust to these factors also need to be investigated. Many practically deployed systems do not record videos at high spatio–temporal resolution in part due to the difficulty in storing the large data that is produced. Hence, dealing with low-resolution video is an important issue. In the approaches discussed so far, it is assumed that reliable features can be extracted in a given setting such as optical flow or background subtracted blobs. In analyzing actions in far-field settings, this assumption does not usually hold. While researchers have addressed these issues in specific settings (cf., [131] and [132]), a systematic and general approach is still lacking. Hence, more research needs to be done to address these practical issues.

B. Invariances in Human Action Analysis

One of the most significant challenges in action recognition is to find methods that can explain and be robust to the wide variability in features that are observed within the same action class. Sheikh *et al.* [133] have identified three important sources that give rise to variability in observed features. They are as follows:

1) viewpoint;
2) execution rate;
3) anthropometry

Any real-world action recognition system needs to be invariant to these factors. In this section, we will review some efforts in this direction that have been pursued in the research community.

1) View Invariance: While it may be easy to build statistical models of simple actions from a single view, it is extremely challenging to generalize them to other views. This is due to the wide variations in motion and structure features induced by camera perspective effects and occlusions. One way to deal with the problem is to store templates from several canonical views as done in [30] and interpolate across the stored views as proposed by [134]. This approach, however, is not scalable because one does not know how many views to consider as canonical. Another approach is to assume that point correspondences across views are available as in [32] and compute a transformation that maps a stored model to an example from an arbitrary view. Similarly, Seitz and Dyer [135] present an approach to recognize cyclic motion that is affine invariant by assuming that feature correspondence between successive time instants is known. It was shown by Rao and Shah [136] that extrema in ST curvature of trajectories are preserved across views, which were exploited to perform view-invariant action recognition. Another example is the work of Parameswaran *et al.* [137] who define a view-invariant representation of actions based on the theory of 2-D and 3-D invariants. In their approach, they consider an action to be a sequence of *poses*. They assume that there exists at least one *key pose* in the sequence in which five points are aligned on a plane in the 3-D world coordinates. Using this assumption, they derive a set of view-invariant descriptors. More recently, the notion of motion history [30] was extended to 3-D by Weinland *et al.* [138] where the authors combine views from multiple cameras to arrive at a 3-D binary occupancy volume. Motion history is computed over these 3-D volumes and view-invariant features are extracted by computing circular fast Fourier transform (FFT) of the volume. All these approaches are strongly tied to the specific choice of feature. There is no general approach of achieving view invariance that can be extended to several features, thus making it an open research issue.

2) Execution Rate Invariance: The second major source of observed variability in features arises from the differences in execution rates while performing the same action. Variations in execution style exist both in interperson and intraperson settings. State–space approaches are robust to minor changes in execution rates, but are not truly rate invariant, because they do not explicitly model transformations of the temporal axis. Mathematically, the variation in execution rate is modeled as a warping function of the temporal scale. The simplest case of linear time warps can be usually dealt with fairly easily. To model highly nonlinear warping functions, the most common method is dynamic time warping (DTW) of the feature sequence such as in [134], [139], and [140]. Recently, Veeraraghavan *et al.* [141] proposed using DTW with constraints to account for the fact that the space of all time-warp functions does not produce physically meaningful actions. DTW is a promising method because it is independent of the choice of feature. The only requirement is

that a distance metric be defined on the feature space. However, DTW requires accurate temporal alignment of test and gallery sequences, i.e., the start and end time instants have to be aligned. Further, the distance computations involved can be prohibitive for long sequences involving many templates. Thus, more efficient methods are required to achieve real-time performance.

3) Anthropometric Invariance: Anthropometric variations such as those induced by the size, shape, gender, etc., of humans is another important class of variabilities that requires careful attention. Unlike viewpoint and execution-rate variabilities that have received significant attention, a systematic study of anthropometric variations has been receiving interest only in recent years. Ad hoc methods that normalize the extracted features to compensate for changes in size, scale, etc., are usually employed when no further information is available. Drawing on studies on human anthropometry Gritai *et al.* [142] suggested that the anthropometric transformation between two different individuals can be modeled as a projective transformation of the image coordinates of body joints. Based on this, they define a similarity metric between actions by using epipolar geometry to provide constraints on actions performed by different individuals. Further research is needed to understand the effects of anthropometric variations and building algorithms to achieve invariance to this factor.

C. Evaluation of Complex Systems

Establishing standardized test beds is a fundamental requirement to compare algorithms and assess progress. It is encouraging to see that several data sets have been made available by research groups and new research is expected to report results on these data sets. Examples include the University of Central Florida (UCF) activity data set [143], Transportation Security Administration (TSA) airport tarmac data set [9], Free Viewpoint National Institute for Research in Computer Science and Control (INRIA) data set [138], and the Royal Institute of Technology (KTH) actions data set [47]. However, most of these data sets consist of simple actions such as opening a closet door, lifting an object, etc. Very few common data sets exist for evaluating higher level complex activities and reasoning algorithms. Complex activity recognition systems consist of a slew of lower level detection and tracking modules. Hence, a straightforward comparison of systems is not easy. One approach to evaluate complex systems is to create ground truth corresponding to outputs from a predefined set of low-level modules. Evaluation would then focus solely on the high-level reasoning engines. While this is one criterion of evaluation, the other criterion is the ability to deal with errors in low-level modules. Participation from the research community is required to address this important issue.

D. Integration With Other Modalities

A vision-based system to recognize human activities can be seen as a crucial stepping stone toward the larger goal of designing machine intelligence systems. To draw a parallel with natural intelligence, humans rely on several modalities including the five classical senses—vision, audition, tactition, olfaction, and gustation—and other senses such as thermoception (temperature) and equilibrioception (balance and acceleration) for everyday tasks. It has also been realized that alternate modalities can improve the performance of vision-based systems, e.g., inertial sensors in structure from motion (SfM), joint audio–video-based tracking [144], etc. Thus, for the longer term pursuit to create machine intelligence, or for the shorter term pursuit of increasing the robustness of action/activity detection modules, integration with other modalities such as audio, temperature, motion, and inertial sensors needs to be investigated in a more systematic manner.

E. Intention Reasoning

Most of the approaches for recognizing and detecting action and activities are based on the premise that the action/activity has already occurred. Reasoning about the intentions of humans and inferring what is going to happen presents a significant intellectual challenge. Security applications are among the first that stand to benefit from such a system, where detection of threat is of utmost importance.

VII. Conclusion

Providing a machine the ability to see and understand as humans do has long fascinated scientists, engineers, and even the common man. Synergistic research efforts in various scientific disciplines, computer vision, artificial intelligence, neuroscience, linguistics, etc., have brought us closer to this goal than at any other point in history. However, several more technical and intellectual challenges need to be tackled before we get there. The advances made so far need to be consolidated, in terms of their robustness to real-world conditions and real-time performance. This would then provide a firmer ground for further research.

References

[1] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception Psychophys.*, vol. 14, no. 2, pp. 201–211, 1973.
[2] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput. Vis. Image Understand.*, vol. 73, no. 3, pp. 428–440, 1999.
[3] C. Cedras and M. Shah, "Motion-based recognition: A survey," *Image Vis. Comput.*, vol. 13, no. 2, pp. 129–155, 1995.
[4] D. M. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Understand.*, vol. 73, no. 1, pp. 82–98, 1999.
[5] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 90–126, 2006.
[6] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, pp. 1–45, 2006.
[7] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with an application to human movement analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1896–1909, Dec. 2005.
[8] B. Georis, M. Maziere, F. Bremond, and M. Thonnat, "A video interpretation platform applied to bank agency monitoring," in *Proc. 2nd Workshop Intell. Distributed Surveillance Syst.*, 2004, pp. 46–50.
[9] N. Vaswani, A. K. Roy-Chowdhury, and R. Chellappa, "Shape activity: A continuous-state HMM for moving/deforming shapes with application to abnormal activity detection," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1603–1616, Oct. 2005.
[10] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The Human ID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.
[11] Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: Current techniques, promising directions and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 4, pp. 39–62, 1999.

[12] S. F. Chang, "The Holy Grail of content-based media analysis," *IEEE Multimedia Mag.*, vol. 9, no. 2, pp. 6–10, Apr. 2002.
[13] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 819–826.
[14] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.
[15] W. Hu, D. Xie, T. Tan, and S. Maybank, "Learning activity patterns using fuzzy self-organizing neural network," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 34, no. 3, pp. 1618–1626, Jun. 2004.
[16] A. Pentland, "Smart rooms, smart clothes," in *Proc. Int. Conf. Pattern Recognit.*, 1998, vol. 2, pp. 949–953.
[17] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational studies of human motion: Part 1, tracking and motion synthesis," *Found. Trends Comput. Graphics Vis.*, vol. 1, no. 2-3, pp. 77–254, 2005.
[18] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–466, 1995.
[19] T. Huang, D. Koller, J. Malik, G. H. Ogasawara, B. Rao, S. J. Russell, and J. Weber, "Automatic symbolic traffic scene analysis using belief networks," in *Proc. Nat. Conf. Artif. Intell.*, 1994, pp. 966–972.
[20] A. K. Roy-Chowdhury and R. Chellappa, "A factorization approach to activity recognition," in *Proc. CVPR Workshop Event Mining*, 2003, p. 41.
[21] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2000, pp. 751–767.
[22] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.
[23] H. Freeman, "On the encoding of arbitrary geometric configurations," *IRE Trans. Electron. Comput.*, vol. 10, no. 2, pp. 260–268, Apr. 1961.
[24] D. G. Kendall, "Shape manifolds, procrustean metrics and complex projective spaces," *Bull. Lond. Math. Soc.*, vol. 16, pp. 81–121, 1984.
[25] H. Blum and R. N. Nagel, "Shape description using weighted symmetric axis features," *Pattern Recognit.*, vol. 10, no. 3, pp. 167–180, 1978.
[26] A. Bissacco, P. Saisan, and S. Soatto, "Gait recognition using dynamic affine invariants," presented at the Int. Symp. Math. Theory Netw. Syst., Leuven, Belgium, Jul. 5–9, 2004.
[27] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2-3, pp. 107–123, 2005.
[28] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveillance Performance Eval. Tracking Surveillance*, 2005, pp. 65–72.
[29] R. Polana and R. C. Nelson, "Detection and recognition of periodic, nonrigid motion," *Int. J. Comput. Vis.*, vol. 23, no. 3, pp. 261–282, 1997.
[30] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.
[31] A. F. Bobick, "Movement, activity, and action: The role of knowledge in the perception of motion," *Philosoph. Trans. Roy. Soc. Lond. B*, vol. 352, pp. 1257–1265, 1997.
[32] T. F. Syeda-Mahmood, M. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," in *Proc. IEEE Workshop Detection Recognit. Events Video*, 2001, pp. 64–72.
[33] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 984–989.
[34] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.
[35] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[36] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2001, pp. 585–591.
[37] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[38] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," *Comput. Vis. Image Understand.*, vol. 73, no. 2, pp. 232–247, 1999.
[39] A. M. Elgammal and C. S. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 681–688.
[40] R. Pless, "Image spaces and video trajectories: Using Isomap to explore video sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1433–1440.
[41] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanism," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 7, no. 5, pp. 923–932, May 1990.
[42] R. A. Young, R. M. Lesperance, and W. W. Meyer, "The Gaussian derivative model for spatial-temporal vision: I. Cortical model," *Spatial Vis.*, vol. 14, no. 3–4, pp. 261–319, 2001.
[43] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[44] O. Chomat and J. L. Crowley, "Probabilistic recognition of activity using local appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1999, vol. 02, pp. 104–109.
[45] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 2, pp. 123–130.
[46] J. C. Niebles, H. Wang, and L. F. Fei, "Unsupervised learning of human action categories using spatial-temporal words," in *Proc. British Mach. Vis. Conf.*, 2006, pp. 1249–1258.
[47] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. Pattern Recognit.*, 2004, pp. 32–36.
[48] S. Savarese, A. Del Pozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlations for unsupervised action classification," presented at the IEEE Workshop Motion Video Comput., Copper Mountain, CO, Jan. 8–9, 2008.
[49] S. Nowozin, G. Bakir, and K. Tsuda, "Discriminative subsequence mining for action classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[50] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17–31, 2007.
[51] S. F. Wong, T. K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.
[52] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 814–827, Jul. 2003.
[53] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[54] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 405–412.
[55] P. A. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[56] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 166–173.
[57] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition," in *Proc. Visual Surveillance Workshop*, 2007, pp. 1–8.
[58] M. A. O. Vasilescu, "Human motion signatures: Analysis, synthesis, recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2002, pp. 456–460.
[59] T. K. Kim, S. F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[60] L. Wolf, H. Jhuang, and T. Hazan, "Modeling appearances with low-rank SVM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.
[61] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[62] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1992, pp. 379–385.

[63] J. Schlenzig, E. Hunter, and R. Jain, "Recursive identification of gesture inputs using hidden Markov models," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, 1994, pp. 187–194.

[64] A. D. Wilson and A. F. Bobick, "Learning visual behavior for gesture analysis," in *Proc. Int. Symp. Comput. Vis.*, 1995, pp. 229–234.

[65] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.

[66] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1163–1173, Sep. 2004.

[67] Z. Liu and S. Sarkar, "Improved gait recognition by gait dynamics normalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 863–876, Jun. 2006.

[68] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 994–999.

[69] D. J. Moore, I. A. Essa, and M. H. Hayes, "Exploiting human actions and object context for recognition tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, pp. 80–86.

[70] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden Markov models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1455–1462.

[71] N. P. Cuntoor and R. Chellappa, "Mixed-state models for nonstationary multiobject activities," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 106–119, 2007.

[72] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto, "Recognition of human gaits," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 2, pp. 52–57.

[73] M. C. Mazzaro, M. Sznaier, and O. Camps, "A model (in) validation approach to gait classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1820–1825, Nov. 2005.

[74] N. P. Cuntoor and R. Chellappa, "Epitomic representation of human activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[75] P. Saisan, G. Doretto, Y. N. Wu, and S. Soatto, "Dynamic texture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 58–63.

[76] A. B. Chan and N. Vasconcelos, "Classifying video with kernel dynamic textures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.

[77] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.

[78] P. V. Overschee and B. D. Moor, "Subspace algorithms for the stochastic identification problem," *Automatica*, vol. 29, no. 3, pp. 649–660, 1993.

[79] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, CRG-TR-96-2, 1996.

[80] K. D. Cock and B. D. Moor, "Subspace angles between ARMA models," *Syst. Control Lett.*, vol. 46, pp. 265–270, 2002.

[81] R. J. Martin, "A metric for ARMA processes," *IEEE Trans. Signal Process.*, vol. 48, no. 4, pp. 1164–1170, Apr. 2000.

[82] R. Vidal and P. Favaro, "Dynamicboost: Boosting time series generated by dynamical systems," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–6.

[83] S. V. N. Vishwanathan, A. J. Smola, and R. Vidal, "Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes," *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 95–119, 2007.

[84] A. Bissacco and S. Soatto, "On the blind classification of time series," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.

[85] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[86] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *Int. J. Comput. Vis.*, vol. 51, no. 2, pp. 91–109, 2003.

[87] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1997, p. 568.

[88] B. North, A. Blake, M. Isard, and J. Rittscher, "Learning and classification of complex dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 1016–1034, Sep. 2000.

[89] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2000, pp. 981–987.

[90] V. Pavlovic and J. M. Rehg, "Impact of dynamic model learning on classification of human motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2000, pp. 1788–1795.

[91] S. M. Oh, J. M. Rehg, T. R. Balch, and F. Dellaert, "Data-driven MCMC for learning and inference in switching linear dynamic systems," in *Proc. Nat. Conf. Artif. Intell.*, 2005, pp. 944–949.

[92] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann, 1988.

[93] H. Buxton and S. Gong, "Visual surveillance in a dynamic and uncertain world," *Artif. Intell.*, vol. 78, no. 1-2, pp. 431–459, 1995.

[94] P. Remagnino, T. Tan, and K. Baker, "Agent orientated annotation in model based visual surveillance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1998, pp. 857–862.

[95] S. Park and J. K. Aggarwal, "Recognition of two-person interactions using a hierarchical Bayesian network," *ACM J. Multimedia Syst.*, vol. 10, Special Issue on Video Surveillance, no. 2, pp. 164–179, 2004.

[96] S. S. Intille and A. F. Bobick, "A framework for recognizing multi-agent action from visual evidence," in *Proc. Nat. Conf. Artif. Intell.*, 1999, pp. 518–525.

[97] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 742–749.

[98] R. L. Kashyap, "Bayesian comparison of different classes of dynamic models using empirical data," *IEEE Trans. Autom. Control*, vol. AC-22, no. 5, pp. 715–727, Oct. 1977.

[99] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[100] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[101] M. I. Jordan, *Learning in Graphical Models*. Cambridge, MA: The MIT Press, 1998.

[102] N. Friedman and D. Koller, "Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks," *Mach. Learn.*, vol. 50, no. 1–2, pp. 95–125, 2003.

[103] C. A. Petri, "Communication with automata," Defense Tech. Inf. Cntr., Fort Belvoir, VA, DTIC Res. Rep. AD0630125, 1966.

[104] R. David and H. Alla, "Petri nets for modeling of dynamic systems a survey," *Automatica*, vol. 30, no. 2, pp. 175–202, 1994.

[105] T. Murata, "Petri nets: Properties, analysis and applications," *Proc. IEEE*, vol. 77, no. 4, pp. 541–580, Apr. 1989.

[106] C. Castel, L. Chaudron, and C. Tessier, "What is going on? a high-level interpretation of a sequence of images," in *Proc. ECCV Workshop Conceptual Descriptions Images*, 1996, pp. 13–27.

[107] N. Ghanem, D. DeMenthon, D. Doermann, and L. Davis, "Representation and recognition of events in surveillance video using Petri nets," in *Proc. 2nd IEEE Workshop Event Mining*, 2004, p. 112.

[108] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. S. Subrahmanian, P. Turaga, and O. Udrea, "A constrained probabilistic petri net framework for human activity detection in video," *IEEE Trans. Multimedia*, to be published.

[109] C. S. Pinhanez and A. F. Bobick, "Human action detection using pnf propagation of temp oral constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1998, p. 898.

[110] Y. Shi, Y. Huang, D. Minen, A. Bobick, and I. Essa, "Propagation networks for recognizing partially ordered sequential action," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 862–869.

[111] Y. Shi, A. F. Bobick, and I. A. Essa, "Learning temporal sequence model from partially labeled data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1631–1638.

[112] R. Hamid, A. Maddi, A. Bobick, and I. Essa, "Structure from statistics—unsupervised activity analysis using suffix trees," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

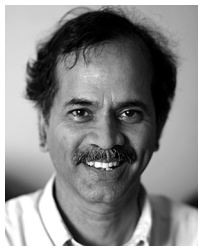
[113] M. Brand, "Understanding manipulation in video," in *Proc. 2nd Int. Conf. Autom. Face Gesture Recognit.*, 1996, p. 94.

[114] M. S. Ryoo and J. K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1709–1718.
[115] J. Earley, "An efficient context-free parsing algorithm," *Commun. ACM*, vol. 13, no. 2, pp. 94–102, 1970.
[116] A. V. Aho and J. D. Ullman, *The Theory of Parsing, Translation, and Compiling, Volume 1: Parsing*. Englewood Cliffs, NJ: Prentice-Hall, 1972.
[117] C. D. L. Higuera, "Current trends in grammatical inference," in *Proc. Joint IAPR Int. Workshops Adv. Pattern Recognit.*, 2000, pp. 28–31.
[118] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 852–872, Aug. 2000.
[119] D. Moore and I. Essa, "Recognizing multitasked activities from video using stochastic context-free grammar," in *Proc. 18th Nat. Conf. Artif. Intell.*, 2002, pp. 770–776.
[120] S. W. Joo and R. Chellappa, "Recognition of multi-object events using attribute grammars," in *Proc. Int. Conf. Image Process.*, 2006, pp. 2897–2900.
[121] N. Rota and M. Thonnat, "Activity recognition from video sequences using declarative models," in *Proc. 14th Eur. Conf. Artif. Intell.*, 2000, pp. 673–680.
[122] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 873–889, Aug. 2001.
[123] S. Hongeng, R. Nevatia, and F. Bremond, "Video-based event recognition: Activity representation and probabilistic recognition methods," *Comput. Vis. Image Understand.*, vol. 96, no. 2, pp. 129–162, 2004.
[124] V. D. Shet, D. Harwood, and L. S. Davis, "Vidmap: Video monitoring of activity with prologue," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 224–229.
[125] S. Tran and L. S. Davis, "Visual event modeling and recognition using Markov logic networks," presented at the IEEE Eur. Conf. Comput. Vis., Marseille, France, Oct. 2008.
[126] D. Chen, J. Yang, and H. D. Wactlar, "Towards automatic analysis of social interaction patterns in a nursing home environment from video," in *Proc. 6th ACM SIGMM Int. Workshop Multimedia Inf. Retrieval*, 2004, pp. 283–290.
[127] A. Hakeem and M. Shah, "Ontology and taxonomy collaborated framework for meeting classification," in *Proc. Int. Conf. Pattern Recognit.*, 2004, pp. 219–222.
[128] S. Guler, J. B. Burns, A. Hakeem, Y. Sheikh, M. Shah, M. Thonnat, F. Bremond, N. Maillot, T. V. Vu, I. Haritaoglu, R. Chellappa, U. Akdemir, and L. Davis, "An ontology of video events in the physical security and surveillance domain," [Online]. Available: http://www.ai.sri.com/~burns/EventOntology, work done as part of the ARDA video event Challenge Workshop, 2003
[129] J. Hobbs, R. Nevatia, and B. Bolles, "An ontology for video event representation," in *Proc. IEEE Workshop Event Detection Recognit.*, 2004, p. 119.
[130] A. R. J. Francois, R. Nevatia, J. Hobbs, and R. C. Bolles, "Verl: An ontology framework for representing and annotating video events," *IEEE MultiMedia Mag.*, vol. 12, no. 4, pp. 76–86, Oct.-Dec. 2005.
[131] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 781–796, Aug. 2000.
[132] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 726–733.
[133] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 144–149.
[134] T. J. Darrell, I. A. Essa, and A. P. Pentland, "Task-specific gesture analysis in real-time using interpolated views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 12, pp. 1236–1242, Dec. 1996.
[135] S. M. Seitz and C. R. Dyer, "View-invariant analysis of cyclic motion," *Int. J. Comput. Vis.*, vol. 25, no. 3, pp. 231–251, 1997.
[136] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 203–226, 2002.
[137] V. Parameswaran and R. Chellappa, "View invariance for human action recognition," *Int. J. Comput. Vis.*, vol. 66, no. 1, 2006.
[138] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 249–257, 2006.
[139] K. Takahashi, S. Seki, E. Kojima, and R. Oka, "Recognition of dexterous manipulations from time-varying images," in *Proc. IEEE Workshop Motion Non-Rigid Articulated Objects*, 1994, pp. 23–28.
[140] M. A. Giese and T. Poggio, "Morphable models for the analysis and synthesis of complex motion patterns," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 59–73, 2000.
[141] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury, "The function space of an activity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 959–968.
[142] A. Gritai, Y. Sheikh, and M. Shah, "On the use of anthropometry in the invariant analysis of human actions," in *Int. Conf. Pattern Recognit.*, 2004, pp. 923–926.
[143] C. Rao and M. Shah, "View-invariance in action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 316–322.
[144] V. Cevher, A. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 715–727, Jun. 2007.

Pavan Turaga (S'05) received the B.Tech. degree in electronics and communication engineering from the Indian Institute of Technology, Guwahati, India, in 2004. He is working towards the Ph.D. degree in electrical engineering at the Department of Electrical and Computer Engineering, University of Maryland, College Park.

He is a recipient of the University of Maryland graduate school fellowship for 2004–2006. His research interests are in statistics and machine learning with applications to computer vision and pattern analysis.

Mr. Turaga is a student member of Association for the Advancement of Artificial Intelligence (AAAI). He was selected to participate in the Emerging Leaders in Multimedia Workshop by IBM, New York, in 2008.



Rama Chellappa (F'92) received the B.E. degree (with honors) in electronics and communication engineering from the University of Madras, Madras, India, in 1975, the M.E. degree (with distinction) in electrical and communication engineering from the Indian Institute of Science, Bangalore, India, in 1977, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1978 and 1981, respectively.

Since 1991, he has been a Professor of Electrical and Computer Engineering and an affiliate Professor of Computer Science at the University of Maryland, College Park. He is also affiliated with the Center for Automation Research (Director), the Institute for Advanced Computer Studies (Permanent Member), the Applied Mathematics program, and the Chemical Physics program. In 2005, he was named a Minta Martin Professor of Engineering. Prior to joining the University of Maryland, he was an Assistant (1981–1986) and Associate Professor (1986–1991) and Director of the Signal and Image Processing Institute (1988–1990) at the University of Southern California (USC), Los Angeles. Over the last 27 years, he has published numerous book chapters and peer-reviewed journal and conference papers. He has also coedited and coauthored many research monographs on Markov random fields, biometrics, and surveillance. His current research interests are in face and gait analysis, 3-D modeling from video, automatic target recognition from stationary and moving platforms, surveillance and monitoring, hyperspectral processing, image understanding, and commercial applications of image processing and understanding.

Dr. Chellappa has served as an Associate Editor of four IEEE TRANSACTIONS. He was a Co-Editor-in-Chief of *Graphical Models and Image Processing*. He also served as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He served as a member of the IEEE Signal Processing Society Board of Governors and as its Vice President of Awards and Membership. Recently, he has been elected to serve a two-year term as the President of the newly constituted IEEE Biometrics Council. He has received several awards, including the National Science Foundation Presidential Young Investigator Award in 1985, four IBM Faculty Development Awards, the 1990 Excellence in Teaching Award from the School of Engineering at USC, the 1992 Best Industry Related Paper Award from the International Association of Pattern Recognition (with Q. Zheng), and the 2000 Technical Achievement Award from IEEE Signal Processing Society. He was elected as

a Distinguished Faculty Research Fellow (1996–1998), and as a Distinguished Scholar Teacher (2003) at the University of Maryland. He coauthored a paper that received the Best Student Paper in the Computer Vision Track at the International Association of Pattern Recognition in 2006. He is a corecipient (with A. Sundaresan) of the 2007 Outstanding Innovator of the Year Award from the Office of Technology Commercialization and received the 2007 A. J. Clark School of Engineering Faculty Outstanding Research Award. He is serving as a Distinguished Lecturer of the IEEE Signal Processing Society for the period 2008–2009 and received the Society's Meritorious Service Award in 2008. He is a Golden Core Member of the IEEE Computer Society, received its Meritorious Service Award in 2004, and has been selected to receive its Technical Achievement Award in 2008. He has served as a General and Technical Program Chair for several IEEE international and national conferences and workshops. He is a Fellow of the International Association for Pattern Recognition.

V. S. Subrahmanian is currently Professor of Computer Science at the University of Maryland, College Park and Director of the University of Maryland's Institute for Advanced Computer Studies (UMIACS). He has worked on nonmonotonic and probabilistic logics, inconsistency management in databases, database models views and inference, rule bases, heterogeneous databases, multimedia databases, probabilistic databases, and agent systems. He has edited two books, one on nonmonotonic reasoning (MIT Press) and one on multimedia databases (Springer-Verlag). He has coauthored an advanced database textbook (Morgan Kaufman, 1997) and a book on heterogeneous software agents. He is the sole author of a textbook on multimedia databases (Morgan Kaufmann).

Prof. Subrahmanian received the NSF Young Investigator Award in 1993 and the Distinguished Young Scientist Award from the Maryland Academy of Science/Maryland Science Center in 1997. He has given invited talks at numerous national and international conferences—in addition, he has served on numerous conference and funding panels, as well as on the program committees of numerous conferences. He has also chaired several conferences. He is or has previously been on the editorial boards of several journals. He has served on DARPA's (Defense Advanced Research Projects Agency) Executive Advisory Council on Advanced Logistics and as an ad hoc member of the U.S. Air Force Science Advisory Board (2001). He also serves on the Board of Directors of the Development Gateway Foundation—an organization that focuses on using information technology in supporting poverty reduction in developing nations.

Octavian Udrea received the B.S. and M.S. degrees from the Polytechnic University of Bucharest, Bucharest, Romania, in 2003 and 2004, respectively, and the Ph.D. degree from the University of Maryland, College Park, in August 2008, all in computer science.

He will join the IBM T. J. Watson Research Center in fall 2008. His primary research interests include knowledge representation, heterogeneous databases, automated code verification, and activity detection in video databases.

Dr. Udrea is a student member of the Association for Computing Machinery (ACM) and the Association for the Advancement of Artificial Intelligence (AAAI).